# Towards Spatial *Supersensing* in Video

**Saining Xie**
**Courant Institute, NYU**
**Oct 2025**

# Benchmarking visual “sensing"

**Benchmarking Difficulty & Bias Risk**

Higher risk of model biases & indirect evaluation

**Visual Question Answering**

*Language* unlock broad querying
— but *introduces bias & shortcut risks*

**From Controlled to Real-World Settings**

Increasing diversity, realism, and task openness

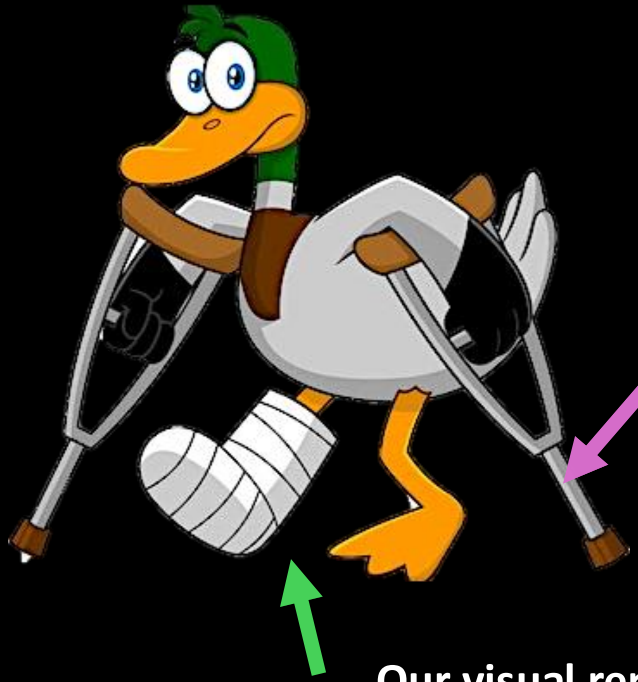
**Structured Vision**

Object/Scene-level reasoning, some ambiguity

**Controlled Tasks**

Fixed classes, clear metrics

Relying too heavily too early on language can act as a shortcut, compensating for the deficiencies in learning effective visual representations.



**Your favorite LLMs**

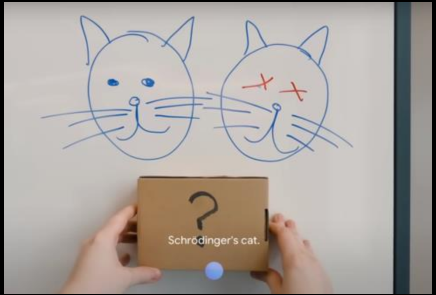
**Our visual representation sucks!**

# Language vs Visual Intelligence

"Who won the game?"

[GPT-4O, OpenAI]

"what does this remind you of?"




[Project Astra, Google]

"Where can I buy this mug?"

[V* - CVPR 2024]

"Which direction leads home?"


[V-IRL - ECCV 2024]

"Thinking in Space"

[TiS - CVPR 2025]

Tasks Requiring more
**Robust Visual-Spatial Intelligence**

Tasks Requiring more
**Strong Language Capability**

# Benchmarking visual “sensing"

**Benchmarking Difficulty & Bias Risk**
Higher risk of model biases & indirect evaluation

**Visual Question Answering**
*Language* unlock broad querying
— but *introduces bias & shortcut risks*

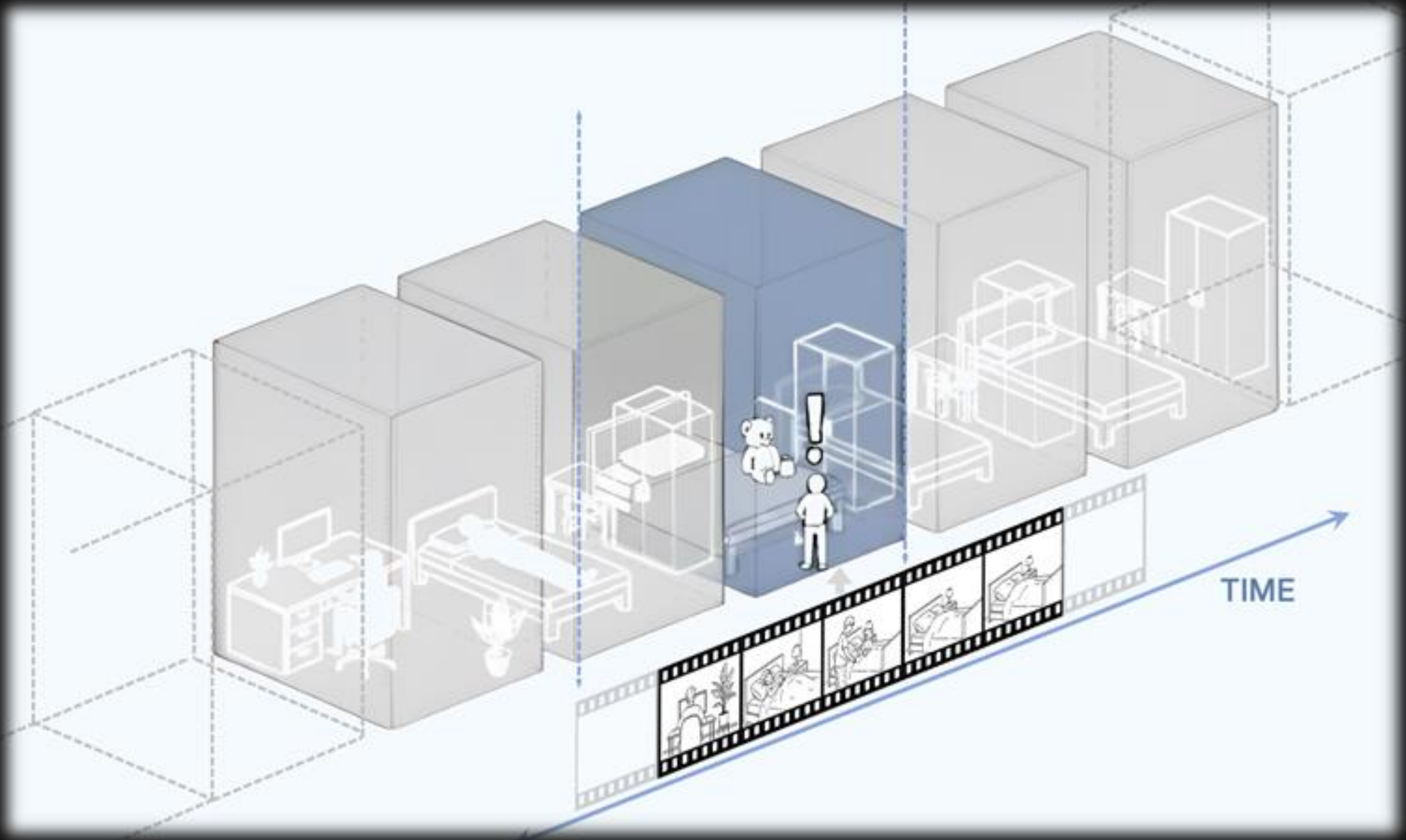
**I think we should really work more on *video* in the multimodal era!**

**From Controlled to Real-World Settings**
Increasing diversity, realism, and task openness

**Structured Vision**
Object/Scene-level reasoning, some ambiguity

**Controlled Tasks**
Fixed classes, clear metrics

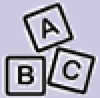
# Why is Video Important?



# Why is Video Important?


**We need to work on "supersensing" for superintelligence!**

# Towards Supersensing in Video



**Linguistic-Only Understanding**

*Knowledge recall;*
*no sensory modeling*

**Linguistic-only understanding:** no multimodal intelligence; reasoning is confined to text and symbols without sensory grounding. Current MLLMs have progressed beyond this stage, yet they still retain traces of its bias.

# Towards Supersensing in Video

**Linguistic-Only Understanding**

*Knowledge recall;*
*no sensory modeling*

**Semantic Perception**

*Naming and describing things for user prompts*

**Semantic perception: parsing pixels into objects, attributes, and relations. This corresponds to the strong "show and tell" capabilities present in MLLMs.**

# Towards Supersensing in Video

**Linguistic-Only Understanding**

*Knowledge recall;*
*no sensory modeling*

**Semantic Perception**

*Naming and describing things for user prompts*

**Streaming Event Cognition**

*Always-on sensing for open-ended streams; memory across time; proactive answering*

**Streaming event cognition: processing live, unbounded streams while proactively interpreting and responding to ongoing events. This aligns with efforts to make MLLMs real-time assistants.**

# Towards Supersensing in Video

**Linguistic-Only Understanding**

*Knowledge recall;*
*no sensory modeling*

**Semantic Perception**

*Naming and describing things for user prompts*

**Streaming Event Cognition**

*Always-on sensing for open-ended streams; memory across time; proactive answering*

**Spatial Cognition**

*Seeing the world behind the video; implicit 3D*

**Implicit 3D spatial cognition: understanding video as projections of a 3D world. Agents must know what is present, where, how things relate, and how configurations change over time. Today's video models remain limited here.**

# Towards Supersensing in Video

**Linguistic-Only Understanding**

*Knowledge recall;*
*no sensory modeling*

**Semantic Perception**

*Naming and describing things for user prompts*

**Streaming Event Cognition**

*Always-on sensing for open-ended streams; memory across time; proactive answering*

**Spatial Cognition**

*Seeing the world behind the video; implicit 3D*

**Predictive World Model**

*Unconscious inference; Predictive, selective, and self-updating world model*

**Predictive world modeling: anticipating future states with an internal model that uses expectation and surprise to organize perception for memory and decision making. This process mirrors human "unconscious inference" and is largely absent in current systems.**

# Towards Supersensing in Video

**Linguistic-Only Understanding**

*Knowledge recall; no sensory modeling*

**Semantic Perception**

*Naming and describing things for user prompts*

**Streaming Event Cognition**

*Always-on sensing for open-ended streams; memory across time; proactive answering*

**Spatial Cognition**

*Seeing the world behind the video; implicit 3D*

**Predictive World Model**

*Unconscious inference; Predictive, selective, and self-updating world model*

TASK-DRIVEN

WORLD MODELING

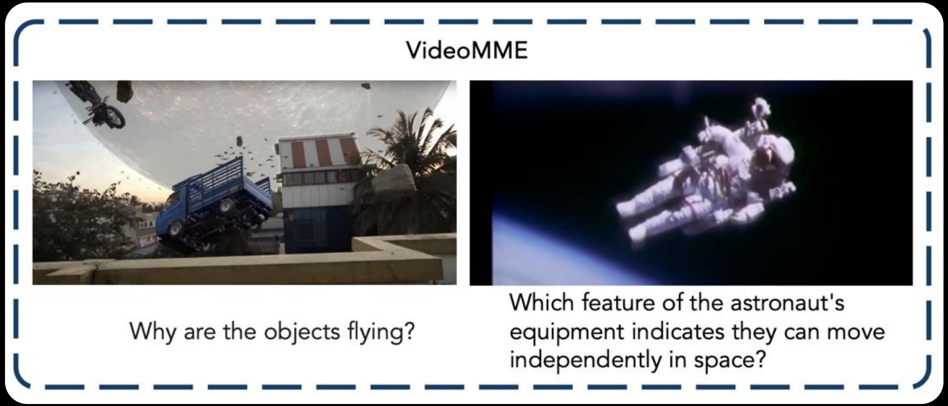
# Current Benchmarks are Not Ready

- Video is the ultimate medium.
  But not all videos are the same.
  Without the right benchmarks, we risk taking the easy path instead of the right one.

| | Previous SoTA | Humans | Gemini 2.5 Flash Preview 04/17* | Gemini 2.5 Pro Preview 05/06* |
|---|---|---|---|---|
| **EVALUATIONS WITH VISUAL INPUTS** | | | | |
| **EgoTempo (test set)** 0-shot open-ended VideoQA | 40.3 (GPT 4.1*) | 63.2 | 36.5 | **43.7** |
| **LVBench (test set)** 0-shot 4-choice VideoQA | 60.1 (GPT 4.1*) | 94.4 | 60.9 | **68.2** |
| **Perception Test (test set)** 0-shot 5-choice VideoQA | 71.4 (Oryx) | 91.4 | 71.2 | **77.3** |
| **QVHighlights (val set)** 4-shot Video Moment Retrieval | **76.1 (Mr BLIP)** | – | 70.2 | 72.6 |
| **VideoMMMU (test set)** 0-shot 5-choice VideoQA | 76.7 (Kimi-k1.6) | 74.4 | 71.9 | **81.3** |
| **1H-VideoQA (test set)** **0-shot 5-choice VideoQA** | 72.2 (Gemini 1.5 Pro) | – | 64.3 | **76.2** |
| **EVALUATIONS WITH AUDIO-VISUAL INPUTS** | | | | |
| **VideoMME (test set, long subset)** 0-shot 4-choice VideoQA | 72.0 (GPT 4.1) | – | 77.8 | **82.0** |
| **YouCook2 Cap (val set)** 4-shot Video Clip Captioning | **198.8 (VAST)** | – | 185.3 | 198.0 |
| **YouCook2 DenseCap (val set)** 4-shot Dense Video Captioning | 67.2 (Vid2Seq) | – | 67.6 | **69.3** |
| **EVALUATIONS WITH VISUAL-SUBTITLES INPUTS** | | | | |
| **Minerva (test set)** 0-shot 5-choice VideoQA | 54.0 (GPT 4.1*) | 92.5 | 61.9 | **63.5** |
| **Neptune (test set)** 0-shot 5-choice VideoQA | 85.1 (GPT 4.1*) | – | 84.5 | **85.4** |
| **EVALUATIONS WITH AUDIO-VISUAL-SUBTITLES INPUTS** | | | | |
| **VideoMME (test set)** 0-shot 4-choice VideoQA | 81.3 (Gemini 1.5 Pro) | – | 79.3 | **85.2** |

**Evaluation of Gemini 2.5 vs. prior models on video understanding benchmarks.**
Performance is measured by string-match accuracy for multiple-choice VideoQA, LLM-based accuracy for EgoTempo, R1@0.5 for QVHighlights and CIDEr for YouCook2. *Videos were processed at 1fps and linearly subsampled to a maximum of 256 frames, except for 1H-VideoQA (7200 frames).

**Some “Spatial Reasoning” benchmark examples**

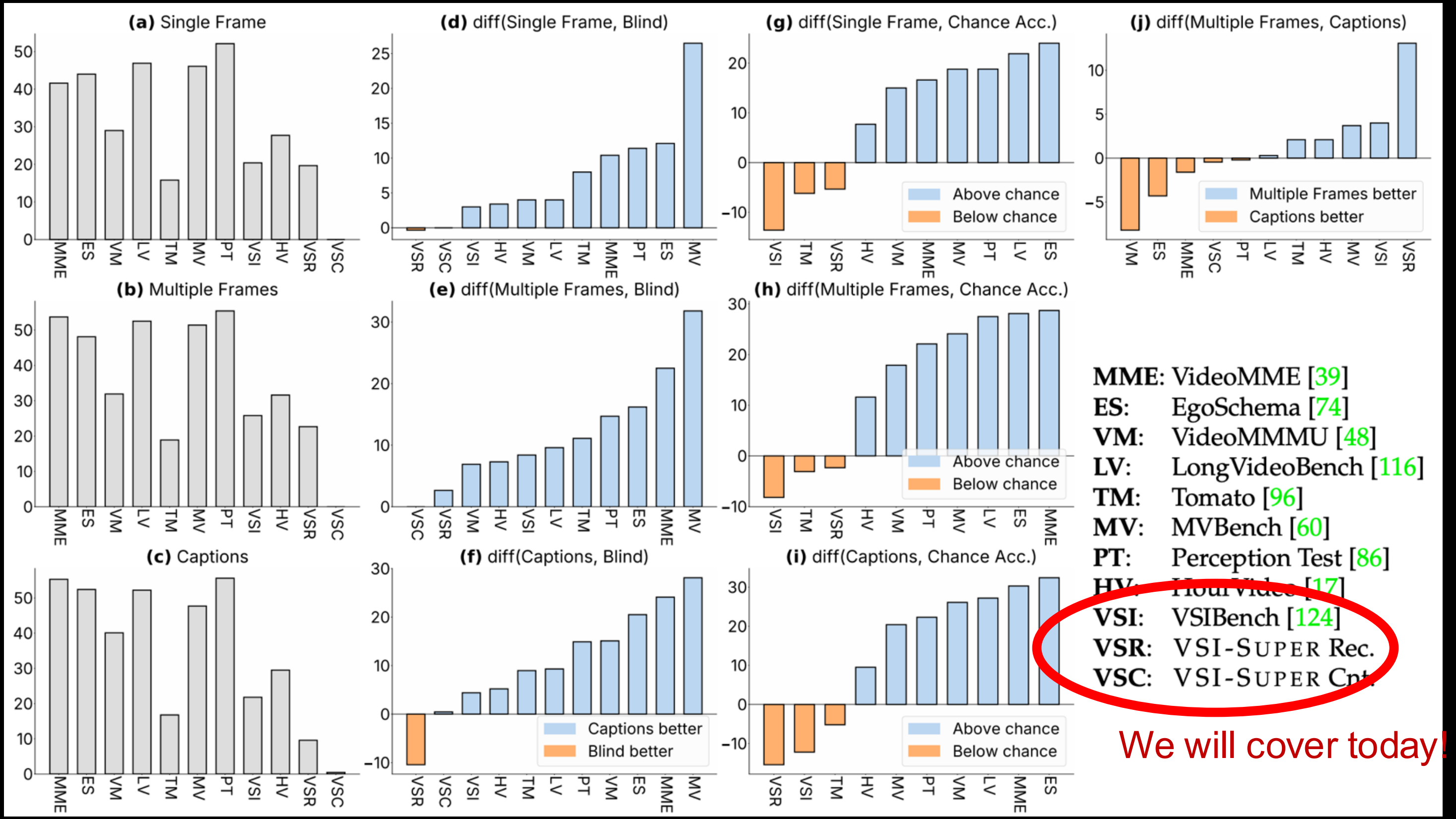



Moravec's Paradox, for video!

# Deconstructing Existing Video Benchmarks

- **Multiple Frames**: Model processes 32 uniformly sampled frames from each video clip — standard video representation method.
- **Single Frame**: Model uses only the middle frame of the clip to test performance with minimal visual context.
- **Frame Captions**: Model receives captions for the same 32 sampled frames (no visual input) to test task solvability without perceptual grounding. Captions generated using the Gemini-2.0-Flash API.

# Deconstructing Existing Video Benchmarks



How can we rigorously investigate spatial supersensing in video, through the creation of new spatial video benchmarks?

# *Thinking in Space:* How Multimodal Large Language Models See, Remember, and Recall Spaces

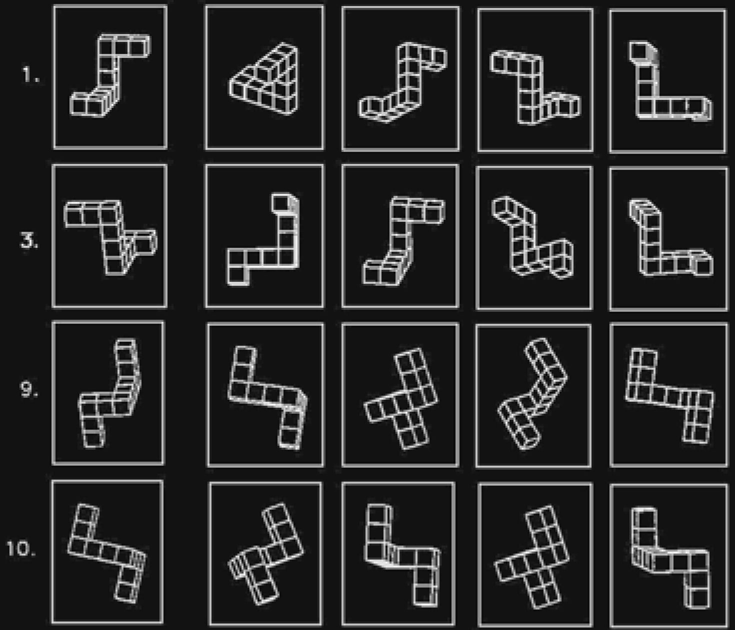
Jihan Yang*, Shusheng Yang*, Anjali W. Gupta*, Rilyn Han*,
Li Fei-Fei, Saining Xie



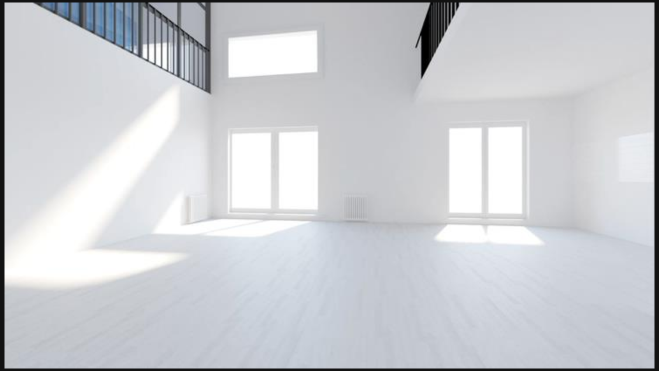
CVPR 2025

# Visual-spatial Intelligence

## Mental Rotation Test



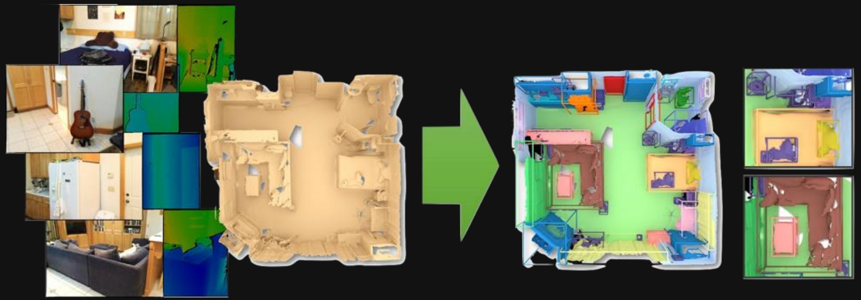

## Furniture Shopping



[1] Shutterstock
[2] Generated with Gemini 2.0 Flash
[3] Adobe Stock
[4] Howard Gardner. Frames of Mind: The Theory of Multiple Intelligences, 1983.

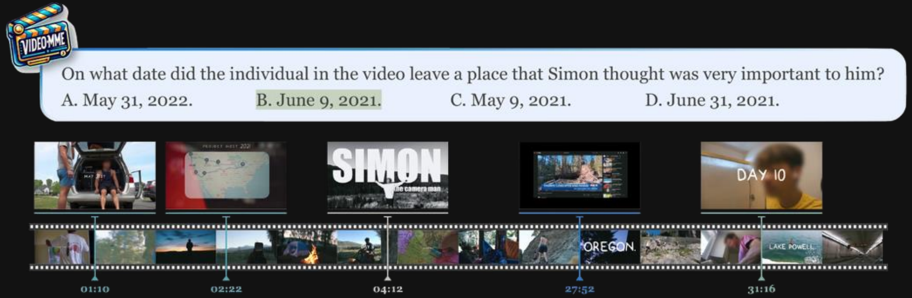
# In computer vision…

We study *space*, but not *thinking*…



[ScanNet, Dai et al. 2017]

We study *thinking*, but not in *space*…




[Video-MME, Fu et al. 2024]

# Watch the video and answer the question

How many chairs are there in this room?

Your Answer: ?

Ground Truth: 9

Gemini-1.5 Pro Answer: 4

# Watch the video and answer the question

If I am standing by the nightstand and facing the chair, is the closet to the left or the right of the chair?

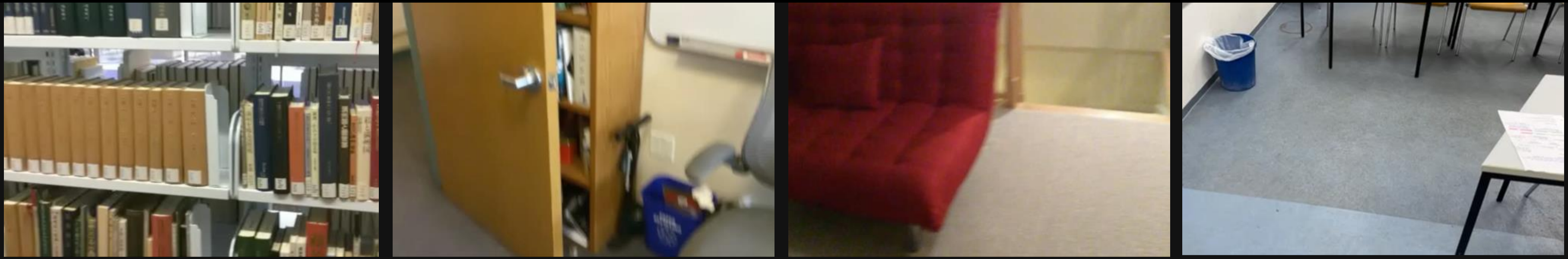
A. Left B. Right

Your Answer: ?

Ground Truth: Left

Gemini-1.5 Pro Answer: Right


How do humans do this?
Can models do this? How?



# *VSI-Bench*

# Benchmark Formulation

Video

Question

If I am standing by the nightstand and facing the chair, is the closet to the left or the right of the chair?

**Model**

GT

Left

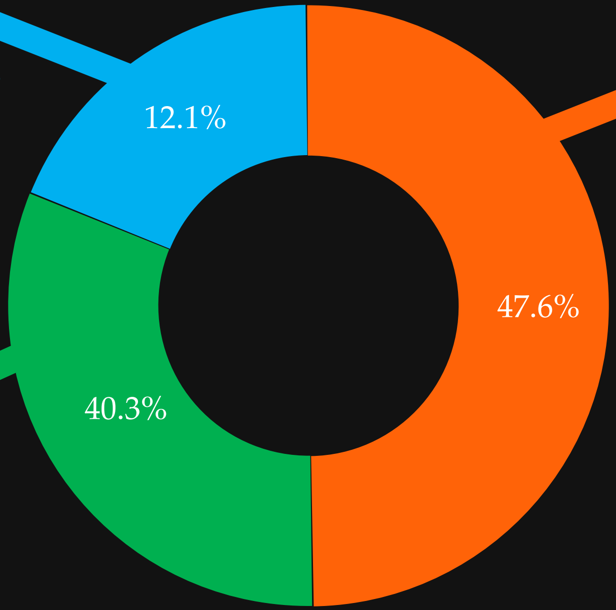
*Difference*

Prediction

# Task Definition

## Spatiotemporal Tasks

*Appearance Order*




## Configuration Tasks

*Object Count*

*Relative Direction*

*Relative Distance*

*Route Plan*

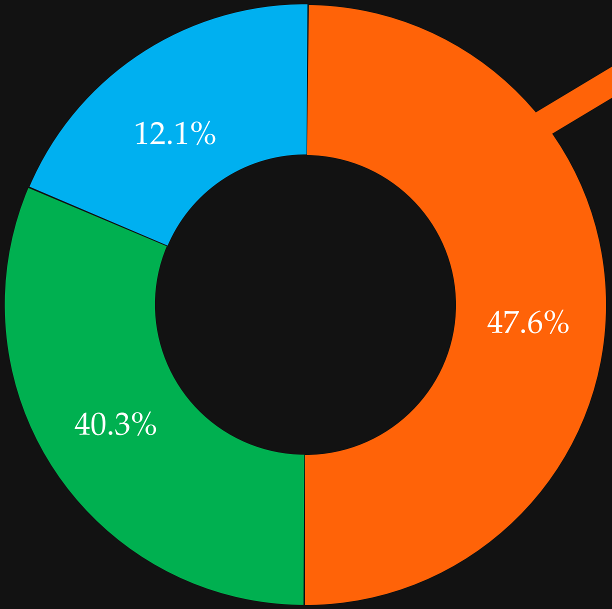
## Measurement Tasks

*Object Size*

*Absolute Distance*

*Room Size*

# Task Definition




## Configuration Tasks

**Object Count**
*How many {object} are there in this room?*

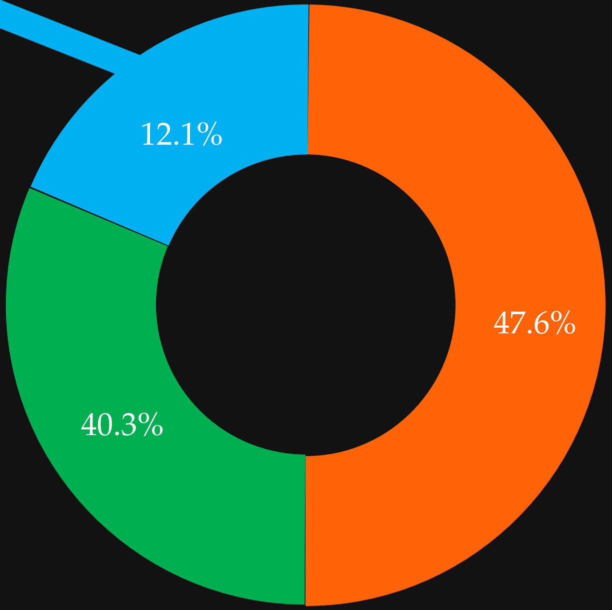
**Relative Direction**
*If I am standing by {object1} and facing {object2}, is {object3} to my left, right, or back?*

**Relative Distance**
*Which of these objects ({list of candidate objects}) is the closest to the {target object}?*

**Route Plan**
*How can I move from {place A} to {place B}? 1. Go forward, 2.______, 3. Go forward, 4. _____.*

# Task Definition

Spatiotemporal Task




**Appearance Order**

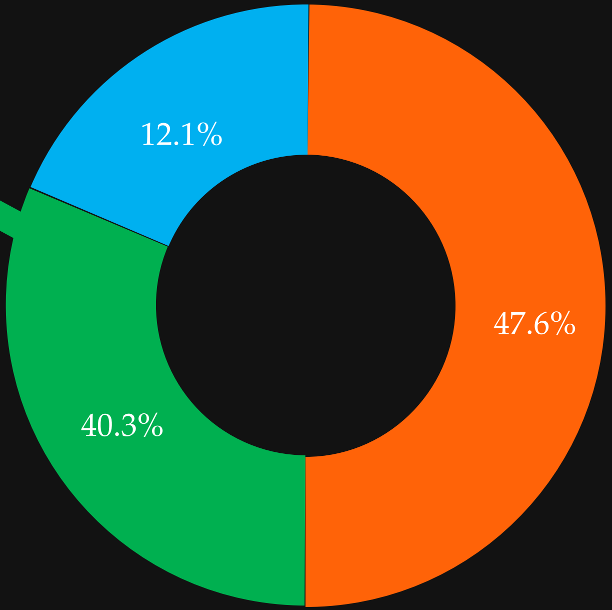
*What is the appearance order of whiteboard, bookshelf, monitor, and cabinet?*

# Task Definition

Measurement Tasks

Object Size

*What is the length of the longest dimension of the whiteboard?*



# How can we construct the benchmark?

Real-world Video

Ground Truth

Object Counts
Room Size
Direction
Distance
...

Build From Scratch

# Repurposing Existing 3D Dataset!

ScanNet
ScanNet++
ARKitScenes

Object Category
3D Boxes
Segmentation Map
...

Object Counts
Object Size
Room Size
Distance
Direction
Appearance
...

Meta Information

Automatic QA Generation

Human In the Loop Verifying and Filtering

5K+ High Quality QA Pairs
Affordable Human Efforts

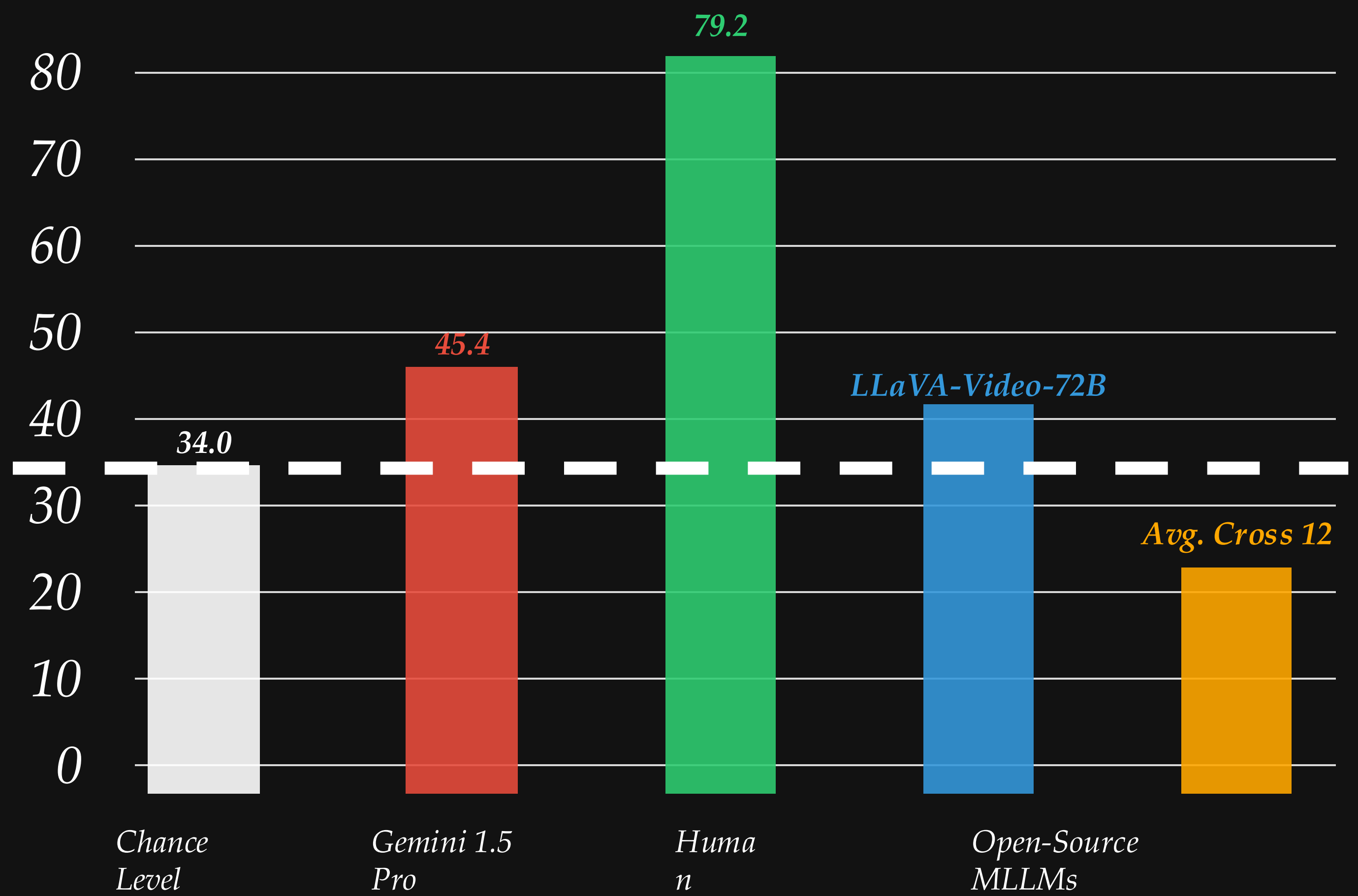
# Benchmarking MLLMs on VSI-Bench

*Chance Level*

*Gemini 1.5 Pro*

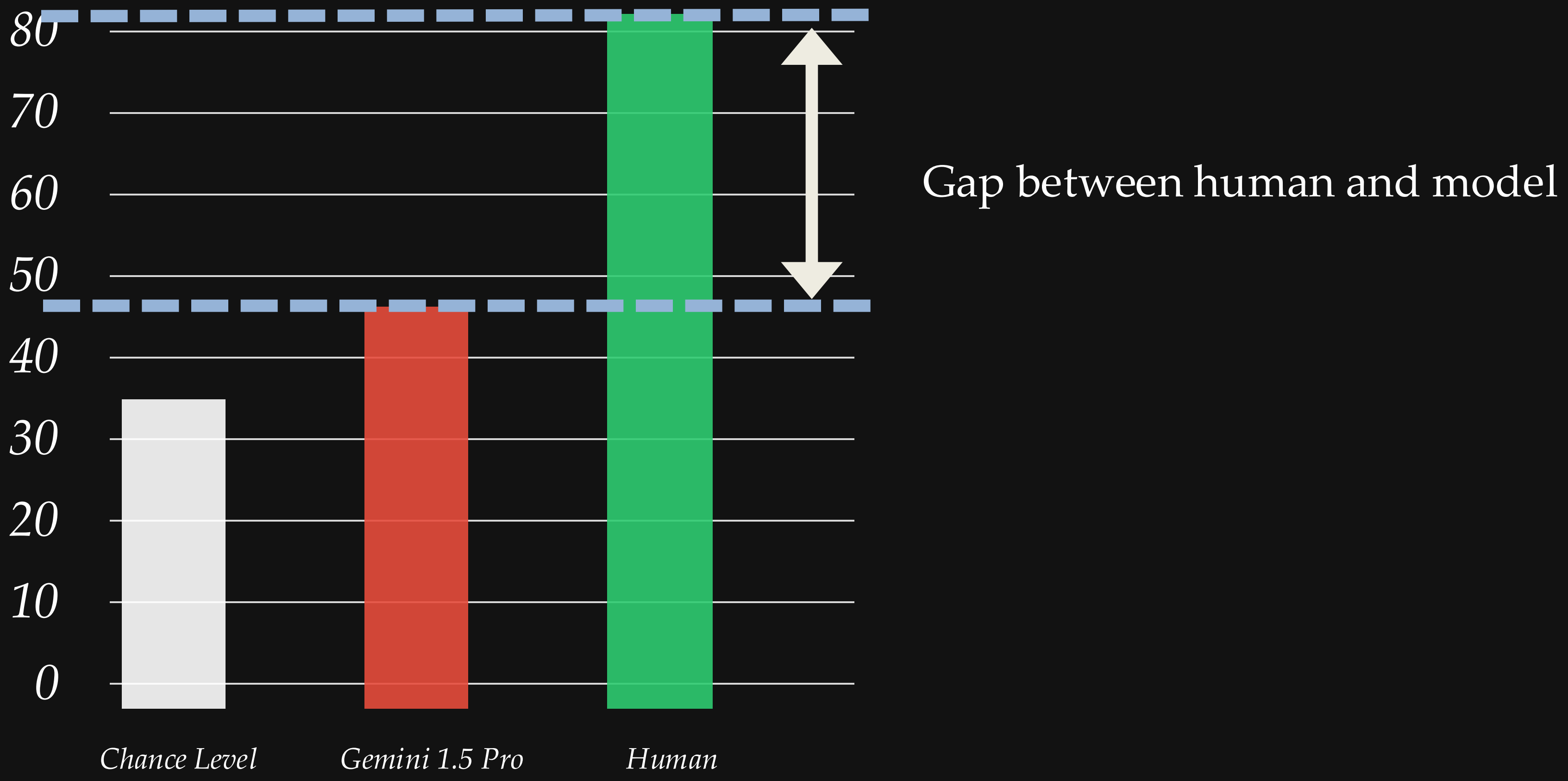
*Huma*
*n*

*Open-Source MLLMs*


79.2
80
70
60
50
45.4
LLaVA-Video-72B
40
34.0
30
Avg. Cross 12
20
10
0
Chance Level
Gemini 1.5 Pro
Huma n
Open-Source MLLMs

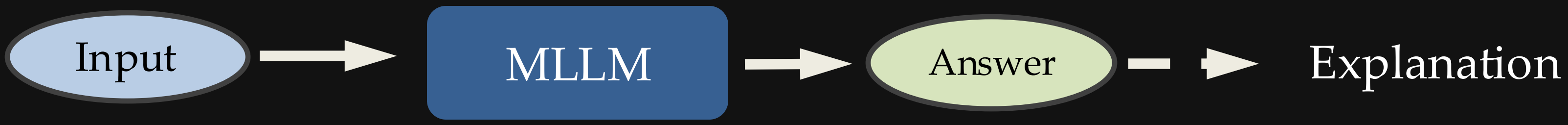
# How do MLLMs Think in Space?

# How do MLLMs Think in Space?



# How do MLLMs Think in Space?

## Prompt Model to Explain Itself



# Analysis by Self-Explanation

# Error Breakdown

Visual Perception

Table's lengths is ....

Recognition error

Linguistic Intelligence

0.36 < 0.30

Logic/Math reasoning error

Relational Reasoning

Spatial Reasoning

Egocentric-Allocentric

The size of telephone is 150 centimeters

Distance/Size/Direction reasoning error

0:13 shows table is on the right of bed

Following the perspective in video instead of the question

## Analysis by Self-Explanation

## Error Breakdown

Visual Perception

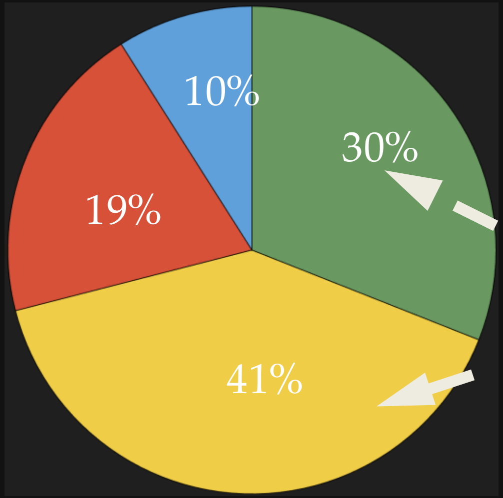
Linguistic Intelligence

Relational Reasoning

Egocentric-Allocentric

Spatial Reasoning

From 163 incorrect samples

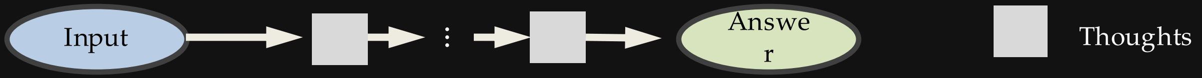


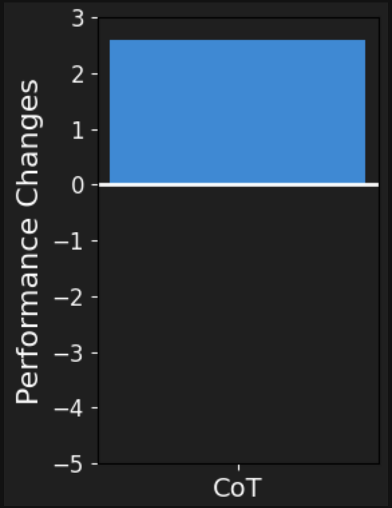

71% spatial reasoning errors

Spatial reasoning is the main bottleneck for MLLMs on VSI-Bench

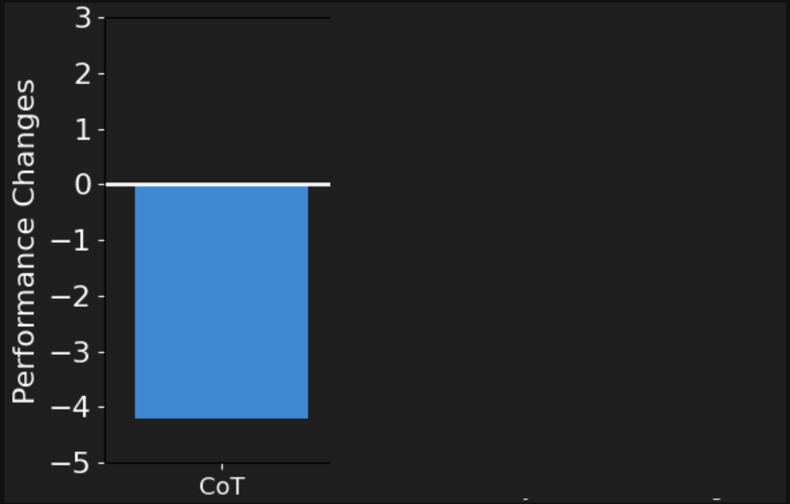
# Scaling Linguistic Reasoning

## Chain-of-thought (CoT)




On Video-MME




On VSI-Bench

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models.

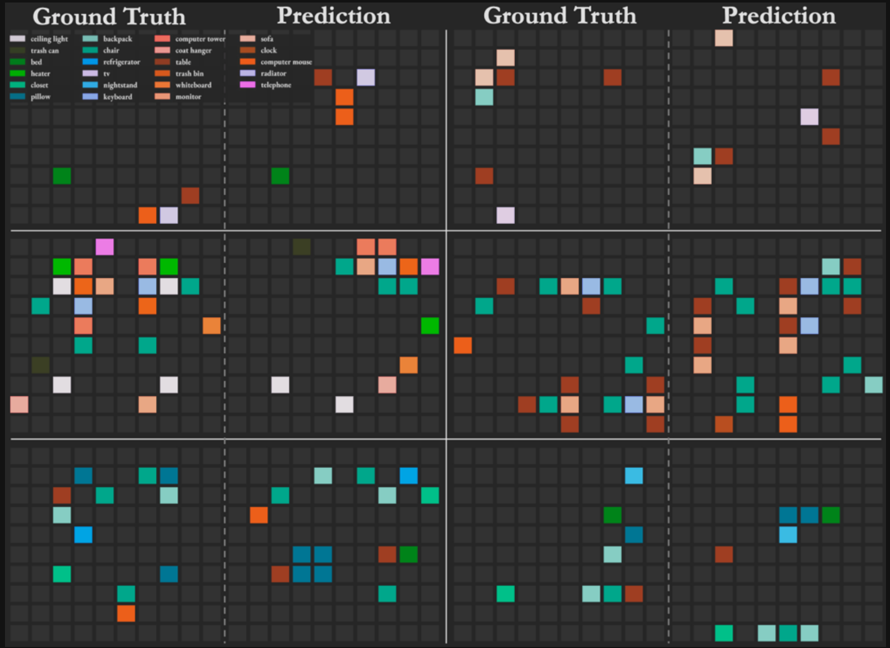
# Analysis by Visualizing Cognitive Map

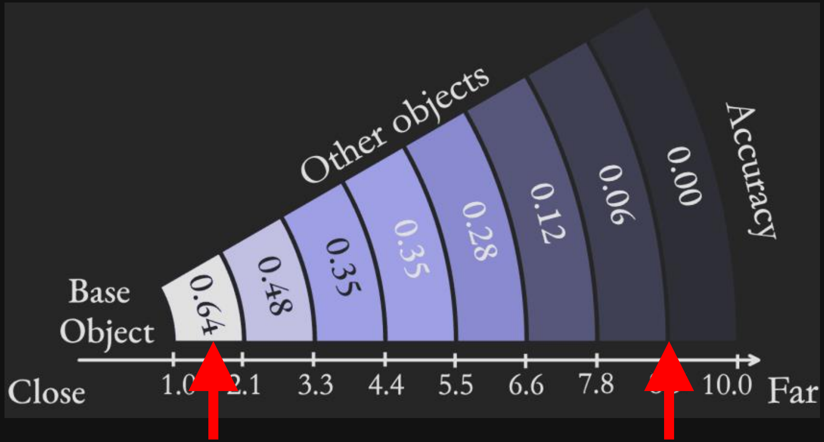
MLLM


10 by 10 cognitive map

Object center positions

# Quantitatively Assess Cognitive Map



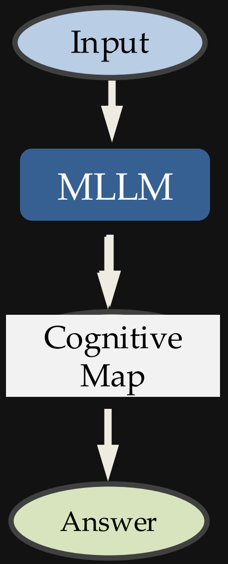



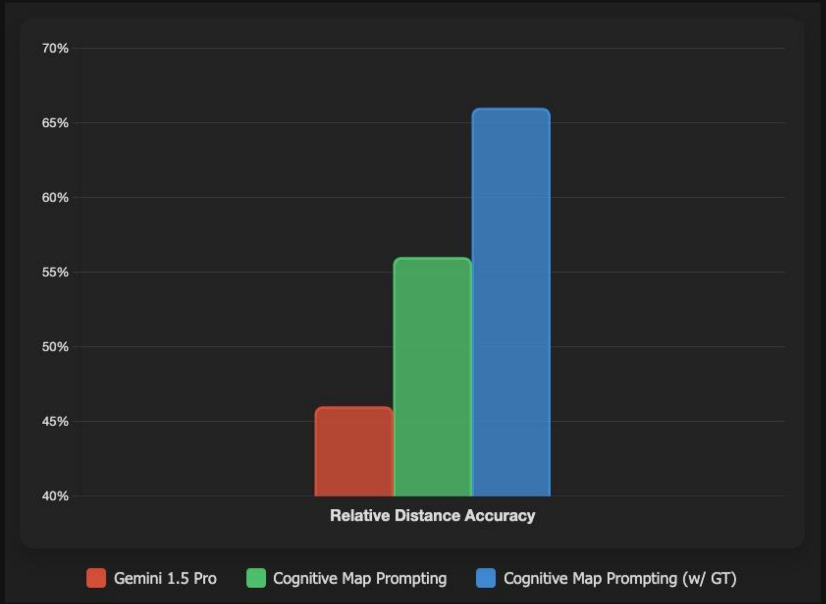

MLLMs are local models

No global understanding

# Can Cognitive Map Help Distance Reasoning?

## Cognitive Map Prompting




Preliminary Step

# *Cambrian-S:*
# Towards Spatial Supersensing in Video

Shusheng Yang[1*] Jihan Yang[1*] Pinzhi Huang[1†] Ellis Brown[1†] Zihao Yang1
Yue Yu[1] Shengbang Tong[1] Zihan Zheng[1] Yifan Xu[1] Muhan Wang[1] Danhao Lu[1]
Rob Fergus[1] Yann LeCun[1] Li Fei-Fei[2] Saining Xie[1]
[1] New York University [2] Stanford University


**Will be on arxiv next week!**

# What’s missing from VSI-Bench:

- **Limited challenge:** videos are typically short in duration.
- **Restricted scope:** confined to a single space.
- **Benchmark-only focus:** lacks exploration of training!

## VSI-SUPER: a two-part, long-horizon evaluation towards evaluating “supersensing”

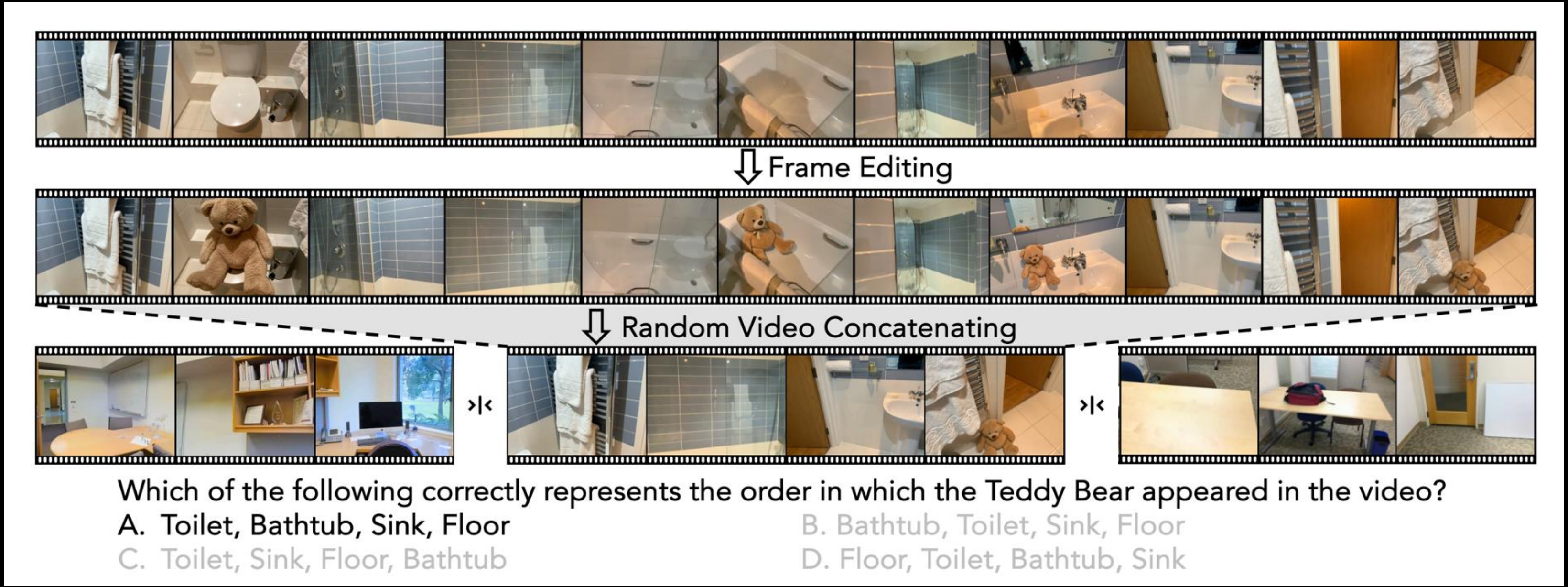
- Combines **concatenated video sequences** with **online Q&A**.

- Like *Needle-in-a-Haystack* tasks but **more realistic and contextually grounded**.

- Designed to be **resistant to brute-force context expansion**, emphasizing true spatial sensing.

# VSI-SUPER Recall:

## Long-horizon spatial observation and recall



Which of the following correctly represents the order in which the Stitch appeared in the video?

A. Stove, Trash bin, Refrigerator, Counter
B. Trash bin, Refrigerator, Counter, Stove
C. Stove, Counter, Refrigerator, Trash bin
D. Trash bin, Stove, Counter, Refrigerator

Which of the following correctly represents the order in which the Hello Kitty appeared in the video?

A. Nightstand, Bed, Crib, Blue bench
B. Blue bench, Crib, Nightstand, Bed
C. Bed, Nightstand, Blue bench, Crib
D. Blue bench, Bed, Crib, Nightstand

Which of the following correctly represents the order in which the Golden Retriever appeared in the video?

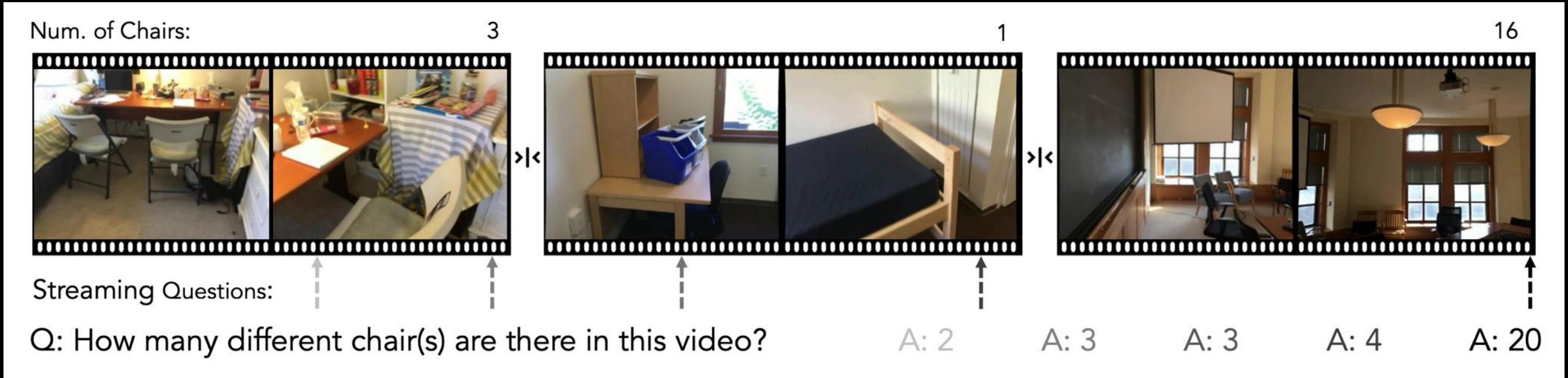
A. Bed, Table, Chest of drawers, Floor
B. Table, Chest of drawers, Bed, Floor
C. Chest of drawers, Floor, Table, Bed
D. Floor, Bed, Chest of drawers, Table

Which of the following correctly represents the order in which the white Ragdoll cat appeared in the video?

A. Ground, Trash bin, Bench, Table
B. Table, Bench, Ground, Trash bin
C. Ground, Trash bin, Table, Bench
D. Trash bin, Bench, Table, Ground

# VSI-SUPER Count:

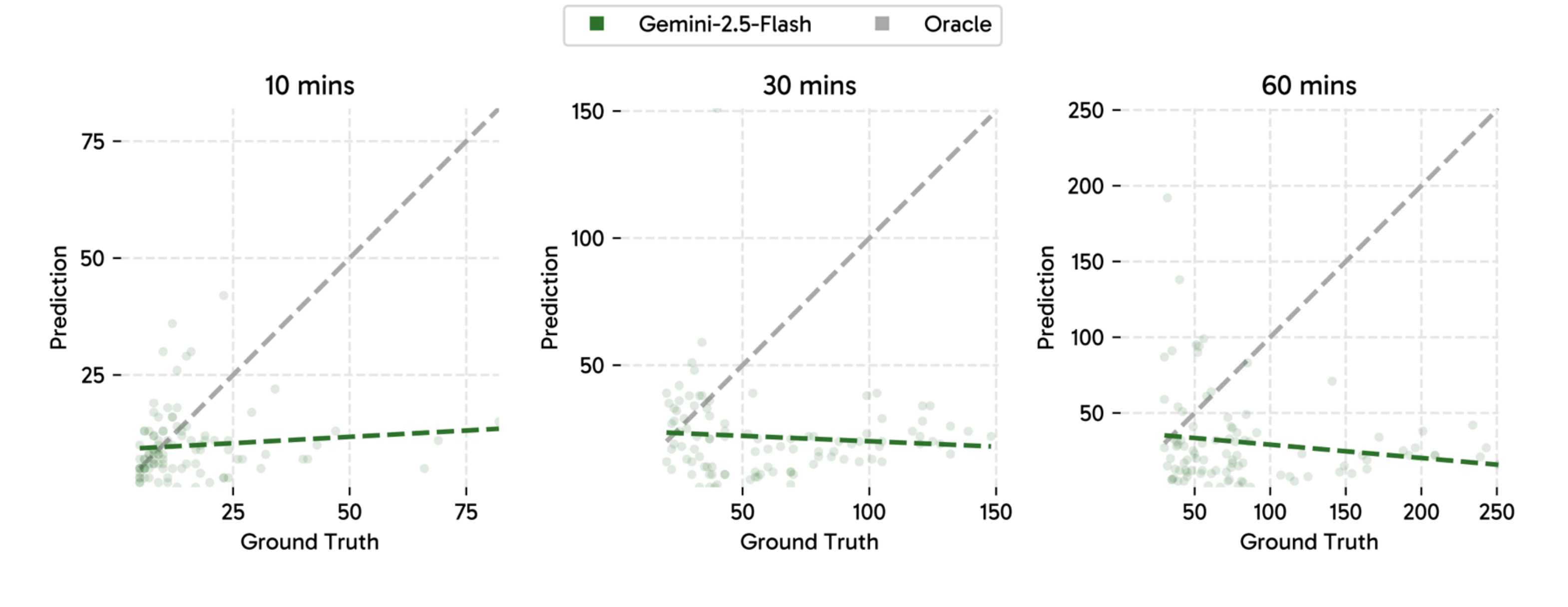
## Continual counting under changing viewpoints and scenes.




Easy for humans, yet extremely difficult for current models!

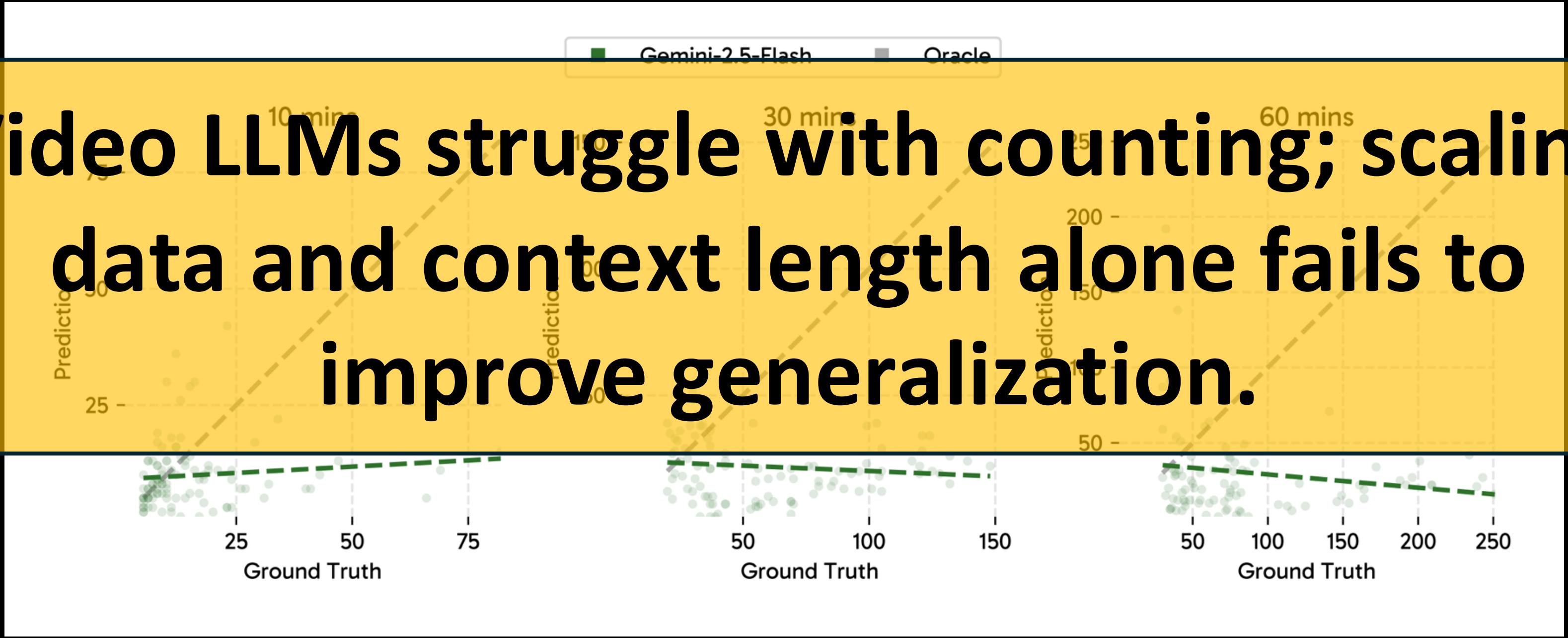
# Gemini-2.5 on VSI-SUPER

| Model | VideoMME[39] | VideoMMMU[48] | VSI-Bench[124] | VSR | | VSC | |
|---|---|---|---|---|---|---|---|
| | | | | 60 mins | 120 mins | 60 mins | 120 mins |
| Gemini-2.5-Flash | 81.5 | 79.2 | 45.7 | 41.5 | Out of Ctx. | 10.9 | Out of Ctx. |



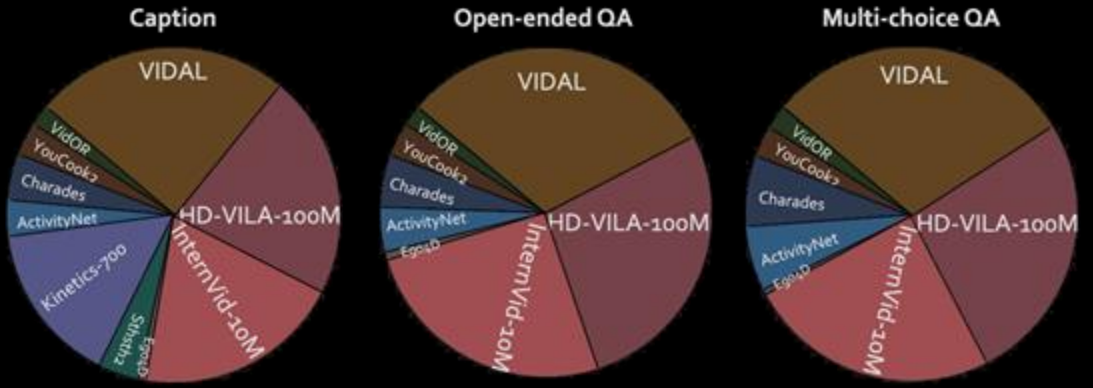
# Gemini-2.5 on VSI-SUPER

| Model | VideoMME[39] | VideoMMMU[48] | VSI-Bench[124] | VSR | | VSC | |
|---|---|---|---|---|---|---|---|
| | | | | 60 mins | 120 mins | 60 mins | 120 mins |
| Gemini-2.5-Flash | 81.5 | 79.2 | 45.7 | 41.5 | Out of Ctx. | 10.9 | Out of Ctx. |




**Video LLMs struggle with counting; scaling data and context length alone fails to improve generalization.**

# Current Data are Not Ready



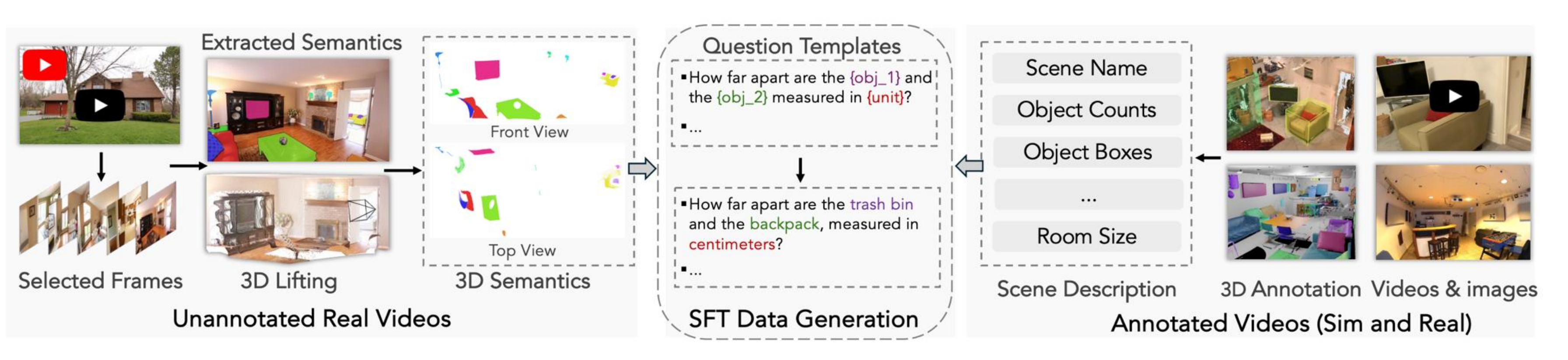

| Dataset | #Caption | #Open-Ended | #Multi-Choice | Dataset | #Caption | #Open-Ended | #Multi-Choice |
|---|---|---|---|---|---|---|---|
| VidOR | 4,018 | 19,875 | 4,773 | Sthsth2 | 8,700 | 0 | 0 |
| YouCook2 | 7,411 | 32,143 | 5,776 | Ego4D | 1,065 | 5,912 | 520 |
| Charades | 9,803 | 48,187 | 13,401 | InternVid-10M | 45,000 | 245,840 | 48,246 |
| ActivityNet | 7,953 | 44,100 | 12,771 | HD-VILA-100M | 48,260 | 263,652 | 51,743 |
| Kinetics-700 | 34,998 | 0 | 0 | VIDAL | 55,000 | 300,472 | 58,968 |

Figure 5: Distribution of data across different datasets and question types (Caption, Open-ended, and Multi-Choice).

**Llava-video-178K!**

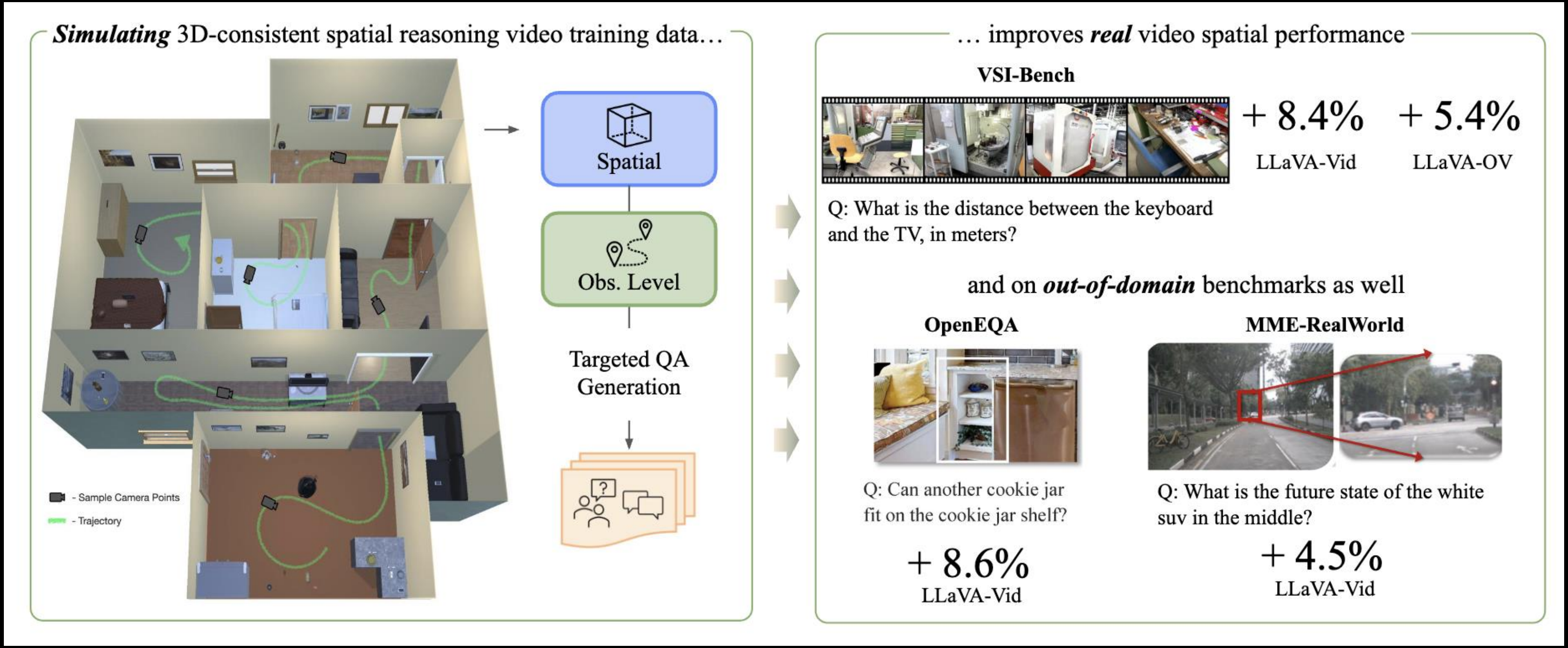
# Current Data are Not Ready

## VSI-590K: Is Spatial Sensing Simply a Data Problem?

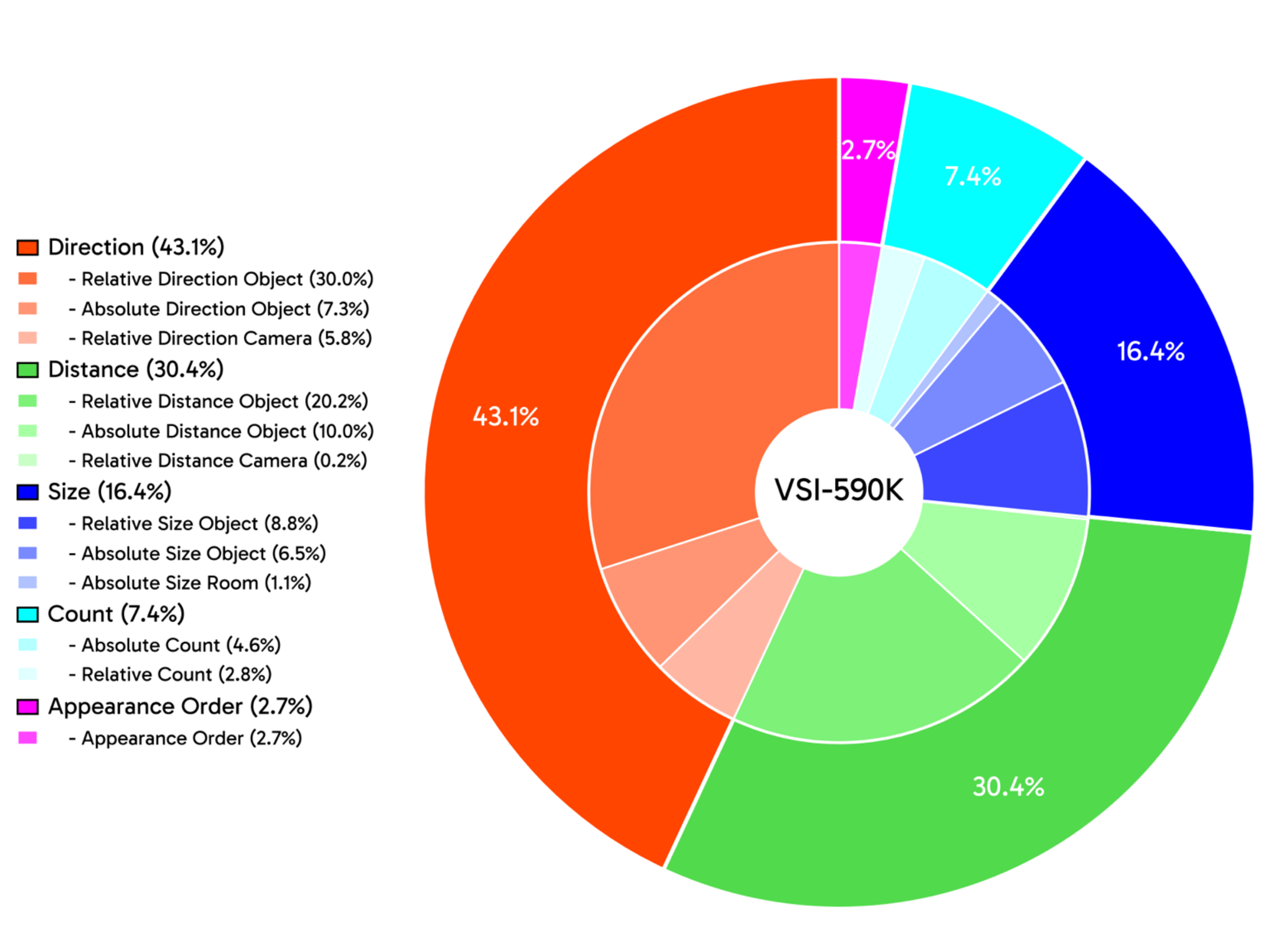


# Data Composition and Sources

| Dataset | # Videos | # Images | # QA Pairs |
|---|---|---|---|
| *Annotated Real Videos* | | | |
| S3DIS [3] | 199 | - | 5,187 |
| Aria Digital Twin [85] | 183 | - | 60,207 |
| ScanNet [31] | 1,201 | - | 92,145 |
| ScanNet++ V2 [129] | 856 | - | 138,701 |
| ARKitScenes [11] | 2,899 | - | 57,816 |
| *Simulated Data* | | | |
| ProcTHOR [34] | 625 | - | 20,092 |
| Hypersim [94] | - | 5,113 | 176,774 |
| *Unannotated Real Videos* | | | |
| YouTube Room Tour | - | 20,100 | 20,100 |
| Open X-Embodiment [83] | - | 14,801 | 14,801 |
| AgiBot-World [15] | - | 4,844 | 4,844 |
| **Total** | 5,963 | 44,858 | 590,667 |




SIMS: Simulated Instruction-Tuning for Multimodal Spatial-Understanding Ellis Brown et al.



Direction (43.1%)
- Relative Direction Object (30.0%)
- Absolute Direction Object (7.3%)
- Relative Direction Camera (5.8%)
Distance (30.4%)
- Relative Distance Object (20.2%)
- Absolute Distance Object (10.0%)
- Relative Distance Camera (0.2%)
Size (16.4%)
- Relative Size Object (8.8%)
- Absolute Size Object (6.5%)
- Absolute Size Room (1.1%)
Count (7.4%)
- Absolute Count (4.6%)
- Relative Count (2.8%)
Appearance Order (2.7%)
- Appearance Order (2.7%)
VSI-590K
43.1%
30.4%
16.4%
7.4%
2.7%

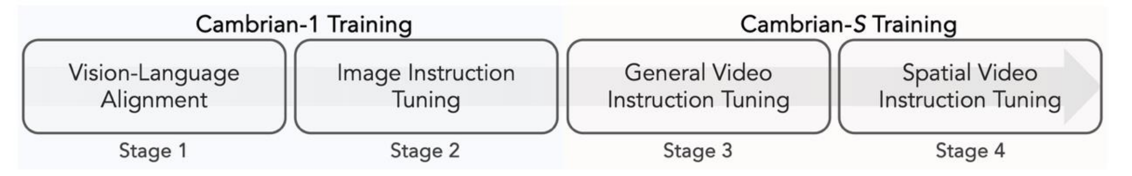
# Data Contributions

| VSI Data Mixture | Image | | | VSI-Bench (Video) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RWQA[1] | 3DSR | CV-B | Avg | Obj Ct | Abs Dst | Obj Sz | Rm Sz | Rel Dst | Rel Dir | Rte Pln | Ap Ord |
| Baseline | 64.2 | 54.5 | 73.5 | 28.5 | 18.1 | 20.0 | 36.0 | 22.2 | 42.9 | 31.3 | 24.6 | 33.0 |
| *Real Videos* | | | | | | | | | | | | |
| + S3DIS | 65.4 | 54.9 | 75.3 | 41.6 | 63.8 | 21.0 | 44.9 | 37.0 | 43.8 | 47.4 | 34.0 | 41.1 |
| + ADT | 65.9 | 56.5 | 77.5 | 41.0 | 51.0 | 29.8 | 52.5 | 40.2 | 42.3 | 38.8 | 34.0 | 39.8 |
| + ARKitScenes | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| + ScanNet | **67.5** | **57.7** | 77.5 | 56.3 | 70.9 | 37.9 | 67.5 | 59.3 | 57.0 | 46.7 | 35.1 | 76.1 |
| + ScanNet++ V2 | 66.1 | 57.3 | 77.5 | 56.3 | 72.5 | 40.7 | 65.7 | 56.9 | 59.7 | 47.1 | 31.4 | 76.2 |
| *Simulated Videos* | | | | | | | | | | | | |
| + ProcThor | 62.2 | 55.7 | 74.9 | 36.4 | 21.0 | 29.7 | 49.3 | 3.8 | 52.3 | 45.7 | 30.4 | 58.7 |
| + HyperSim | 67.2 | 56.0 | **79.7** | 45.6 | 67.8 | 32.0 | 59.3 | 36.4 | 53.2 | 47.0 | 32.5 | 36.6 |
| *Pseudo-Annotated Images* | | | | | | | | | | | | |
| + YTB RoomTour | 62.2 | 52.6 | 75.0 | 32.5 | 43.4 | 25.8 | 24.2 | 27.3 | 38.7 | 31.4 | 28.4 | 40.9 |
| + OXE & AGIBot | 64.4 | 54.4 | 72.5 | 30.6 | 40.3 | 23.1 | 27.9 | 26.6 | 38.0 | 22.8 | 32.0 | 33.8 |
| **All-in-One** | 60.8 | 54.0 | 77.9 | **63.2** | **73.5** | **49.4** | **71.4** | **70.1** | **66.9** | **61.5** | 36.6 | **76.6** |

**Real and synthetic data together provide rich sources that boost spatial understanding.**

# Pre-Training is Important

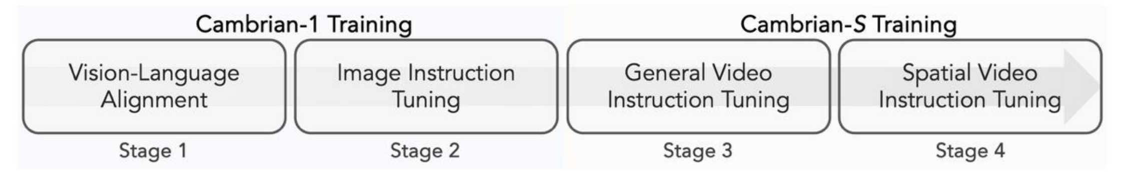
| Model | VSI-Bench | VideoMME | EgoSchema | Perception Test |
|---|---|---|---|---|
| Different Base Models | | | | |
| A1 (*w/o.* I-IT, *i.e.* QwenLM) | 21.4 | 44.2 | 42.9 | 44.5 |
| A2 (A1 + I-IT, *i.e.* Cambrian-1) | 25.8 | 53.7 | 48.1 | 55.4 |
| A3 (A2 + V-IT, 429K data) | 28.9 | 61.2 | 50.3 | 66.3 |
| A4 (A2 + V-IT, 3M data) | **35.7** | **62.6** | **77.0** | **70.9** |
| SFT *w/.* VSI-590K | | | | |
| from A1 | 57.2 | 40.3 | 38.7 | 52.3 |
| from A2 | 66.8 | 46.7 | 47.2 | 52.3 |
| from A3 | 68.8 | 52.3 | 48.4 | 55.8 |
| from A4 | **69.2** | **54.1** | **55.2** | **59.2** |
| SFT *w/.* VSI-590K & general V-IT data mixture | | | | |
| from A1 | 61.3 | 60.5 | 52.8 | 65.0 |
| from A2 | 63.2 | **62.6** | 52.9 | 65.6 |
| from A3 | 64.0 | 61.0 | 54.9 | 66.8 |
| from A4 | **65.1** | 61.9 | **77.3** | **71.2** |



# Pre-Training is Important

| Model | VSI-Bench | VideoMME | EgoSchema | Perception Test |
|---|---|---|---|---|
| Different Base Models | | | | |
| A1 (*w/o.* I-IT, *i.e.* QwenLM) | 21.4 | 44.2 | 42.9 | 44.5 |
| A2 (A1 + I-IT, *i.e.* Cambrian-1) | 25.8 | 53.7 | 48.1 | 55.4 |
| A3 (A2 + V-IT, 429K data) | 28.9 | 61.2 | 50.3 | 66.3 |
| A4 (A2 + V-IT, 3M data) | **35.7** | **62.6** | **77.0** | **70.9** |
| SFT *w/.* VSI-590K | | | | |
| from A1 | 57.2 | 40.3 | 38.7 | 52.3 |
| from A2 | 66.8 | 46.7 | 47.2 | 52.3 |
| from A3 | [illegible] | [illegible] | 48.4 | [illegible] |
| from A4 | [illegible] | [illegible] | [illegible] | [illegible] |
| SFT *w/.* VSI-590K & general V-IT data mixture | | | | |
| from A1 | [illegible] | [illegible] | 52.8 | [illegible] |
| from A2 | [illegible] | [illegible] | [illegible] | [illegible] |
| from A3 | 64.0 | 61.0 | 54.9 | 66.8 |
| from A4 | **65.1** | 61.9 | **77.3** | **71.2** |

**The quality of *multimodal* pre-training strongly influences post-training effectiveness.**



**Current architectures are not ready**

What makes spatial sensing unique?

- Infinite tokens in, infinite tokens out

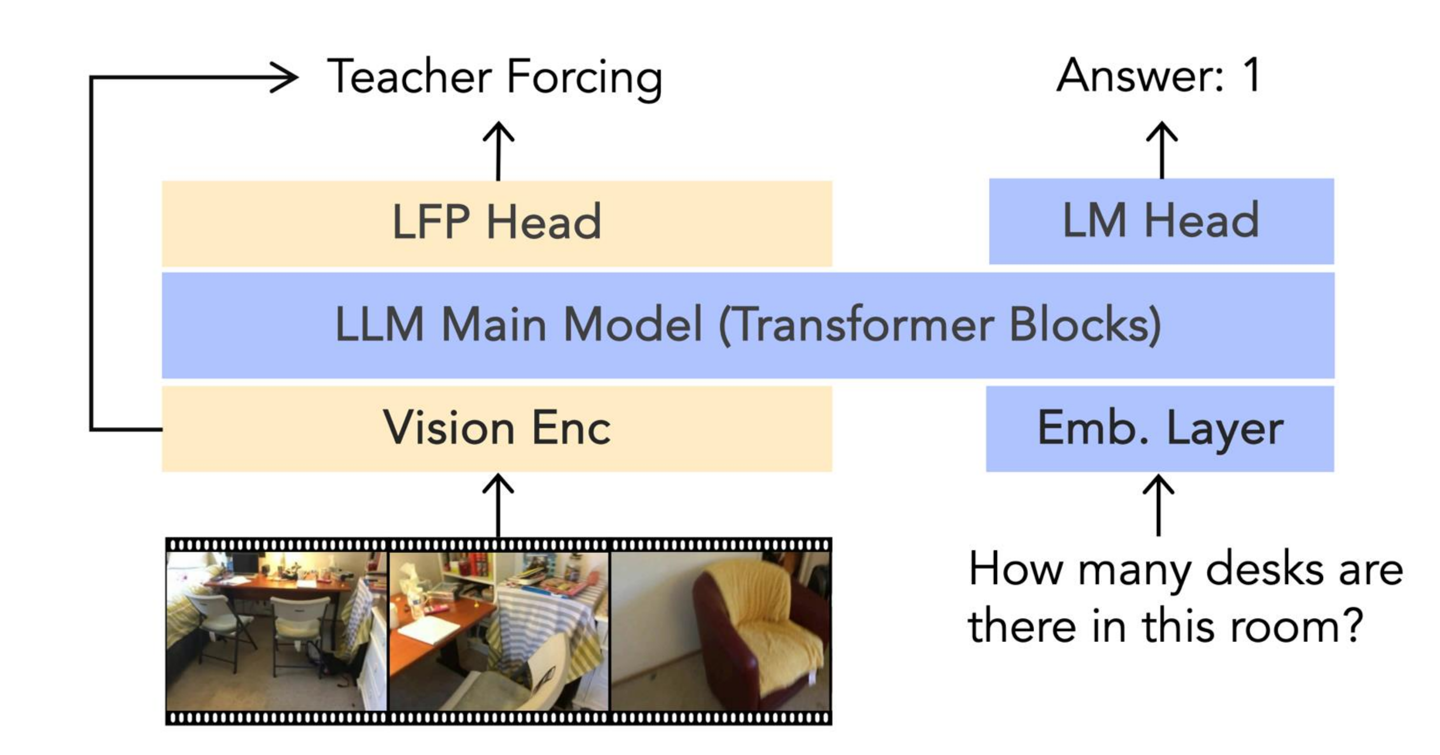
- Our real-world experience isn’t meant to be processed token by token.

# Current architectures are not ready

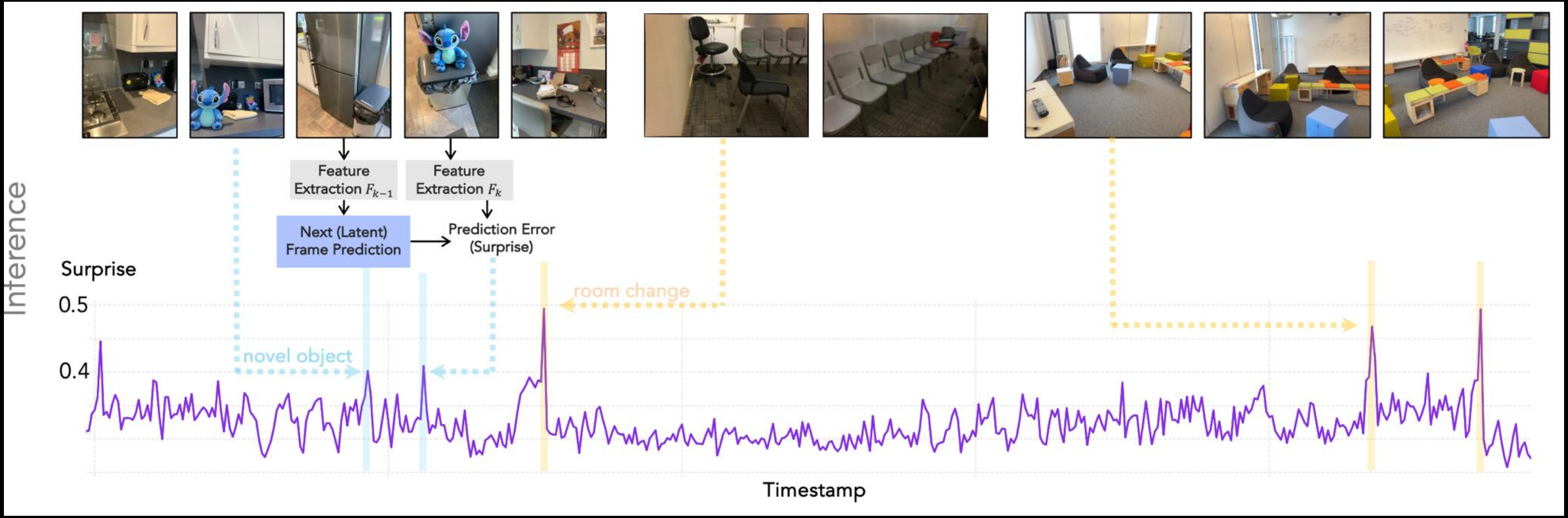
- 🧠 **Human Visual Stream = Extremely High Bandwidth**
- 👁️ **Retina → Brain:** ~10 **million bits/sec**
- 🌐 **All sensory input (mostly vision):** up to **1 billion bits/sec**
- 🧩 **Conscious awareness:** only **~10 bits/sec**

**Most visual data is filtered and compressed before reaching perception. How?**

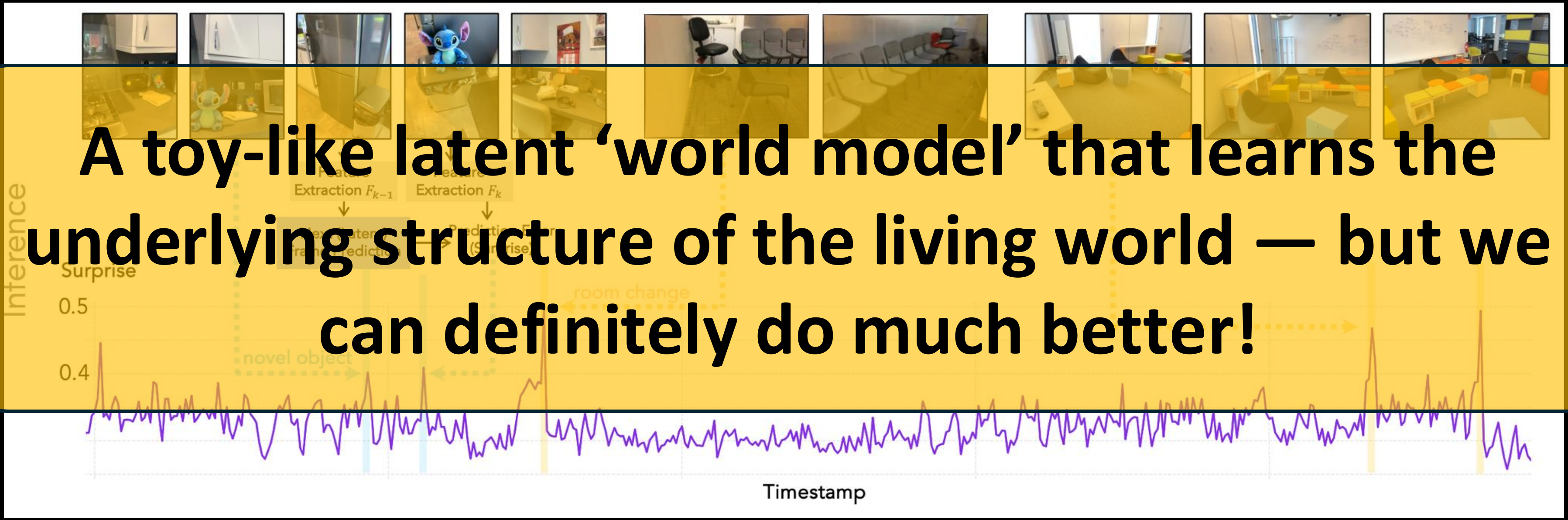
# Prototype: Predictive Sensing



# **Violation-of-Expectation (or simply, surprises!):**
how humans regulate what information they take in.

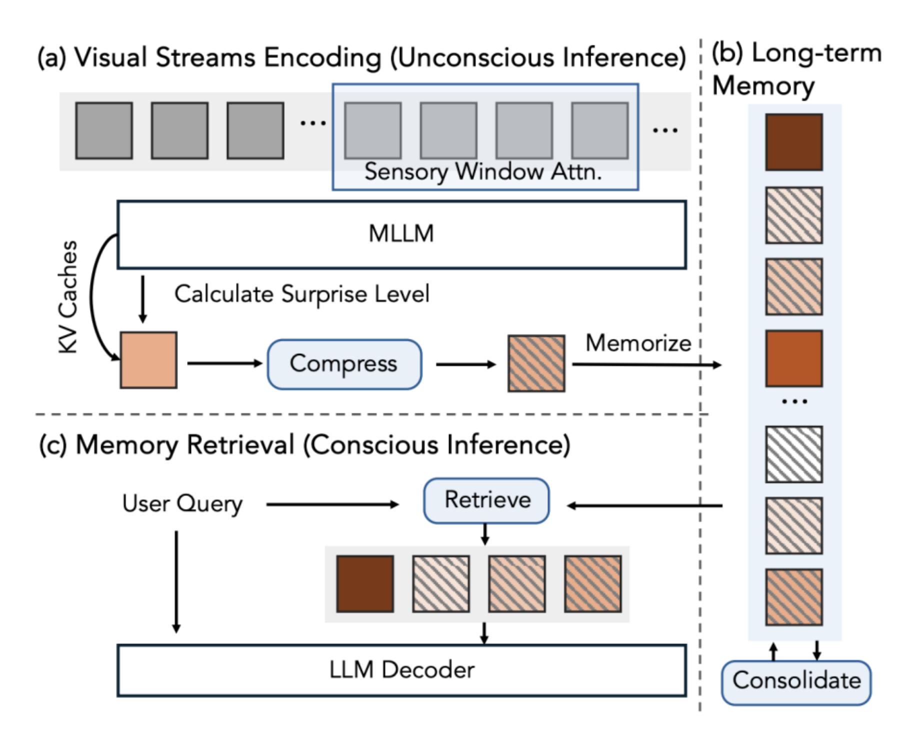


# Violation-of-Expectation (or simply, surprises!):
how humans regulate what information they take in.

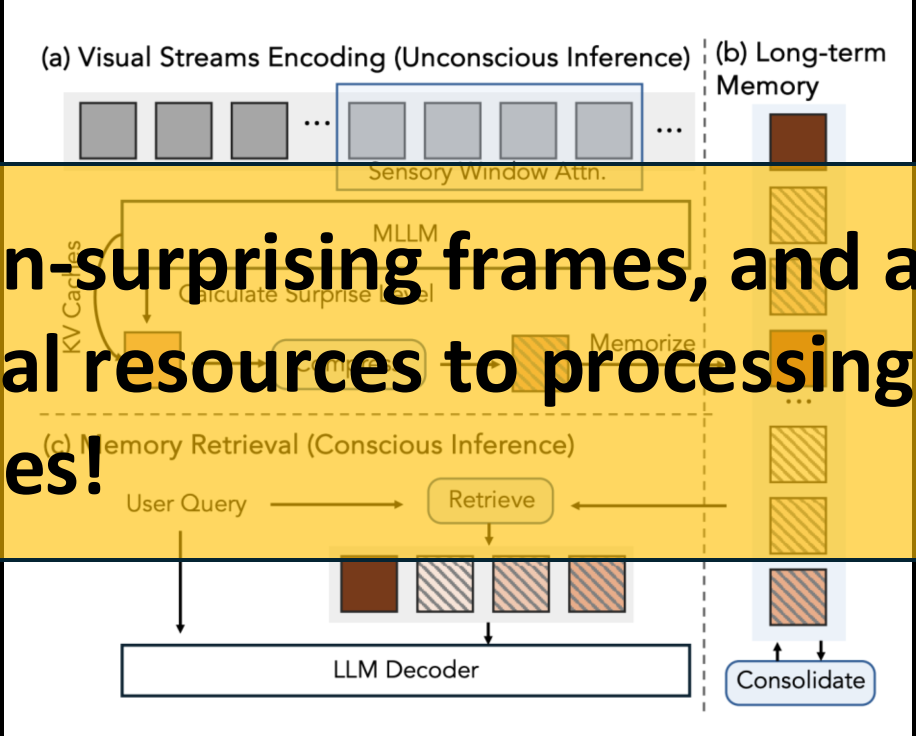



**A toy-like latent 'world model' that learns the underlying structure of the living world — but we can definitely do much better!**

# Use Case #1: Memory Management

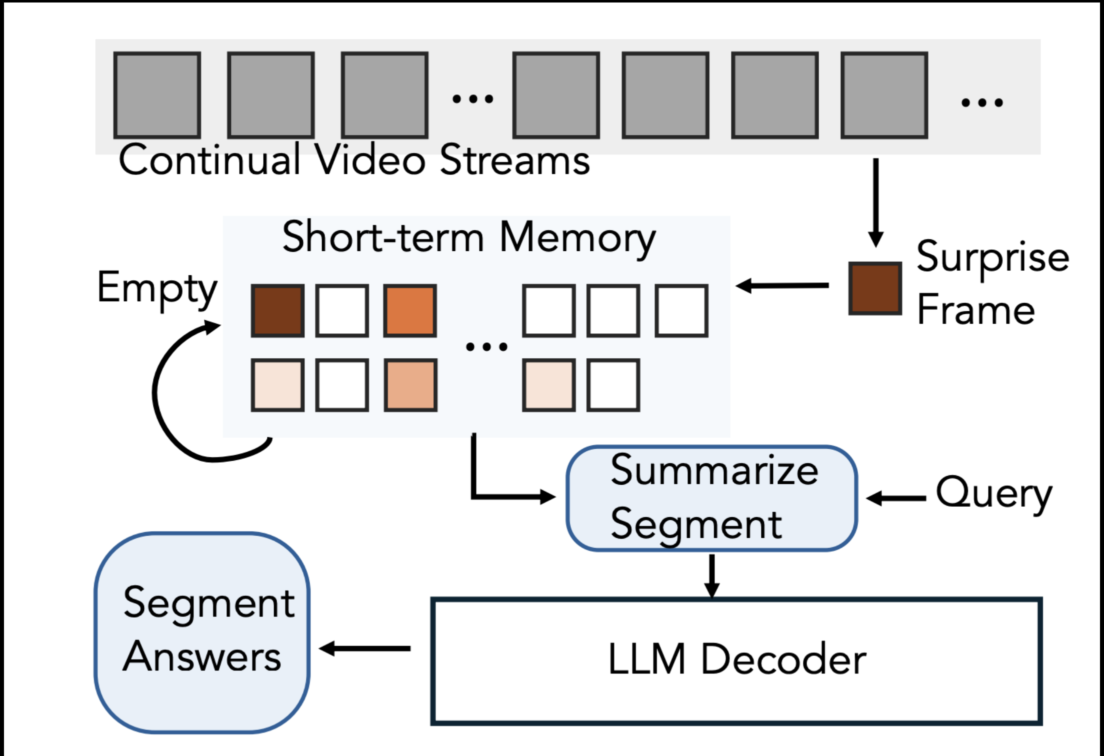


# Use Case #1: Memory Management

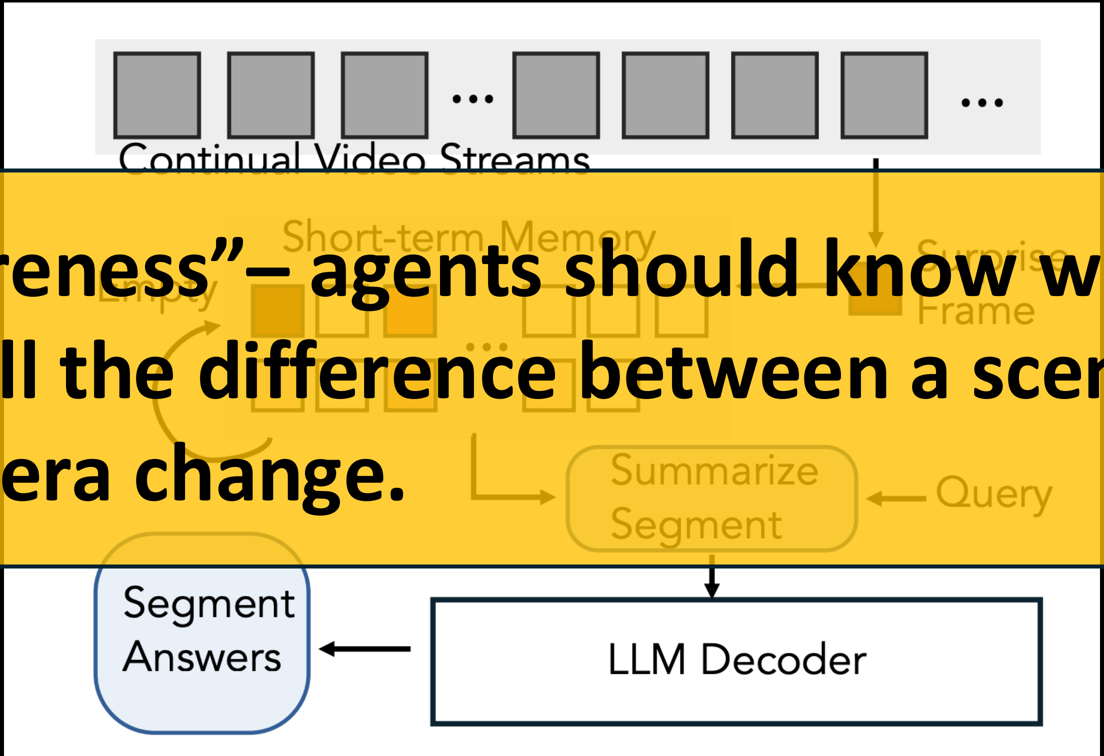



**Compress non-surprising frames, and allocate more computational resources to processing and storing surprising ones!**

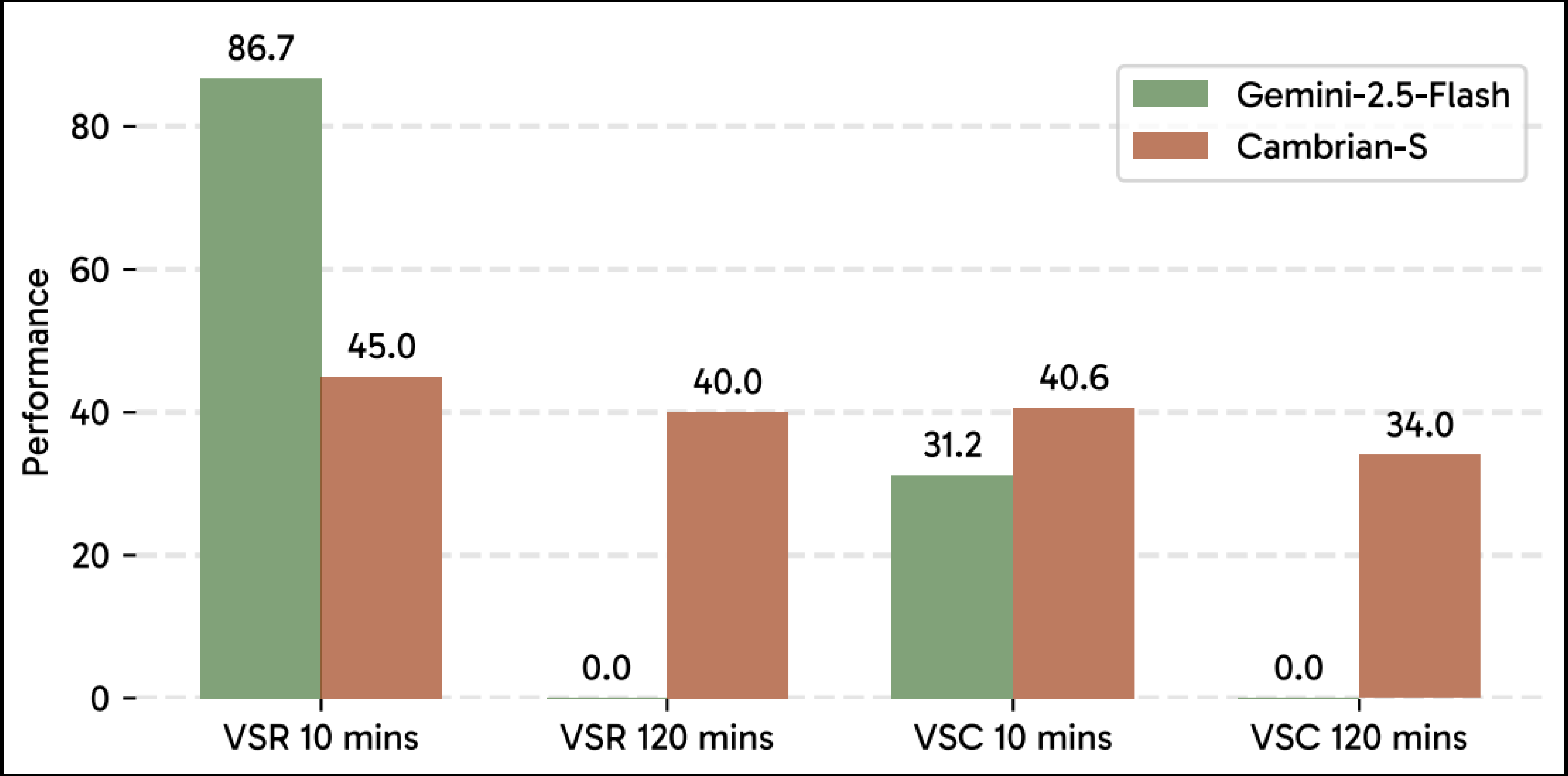
# Use Case #2: Scene/Event Segmentation



# Use Case #2: Scene/Event Segmentation




**"Self-awareness"– agents should know where they are and tell the difference between a scene change and a camera change.**



86.7
45.0
40.0
40.6
31.2
34.0
0.0
0.0
Gemini-2.5-Flash
Cambrian-S
Performance
0
20
40
60
80
VSR 10 mins
VSR 120 mins
VSC 10 mins
VSC 120 mins

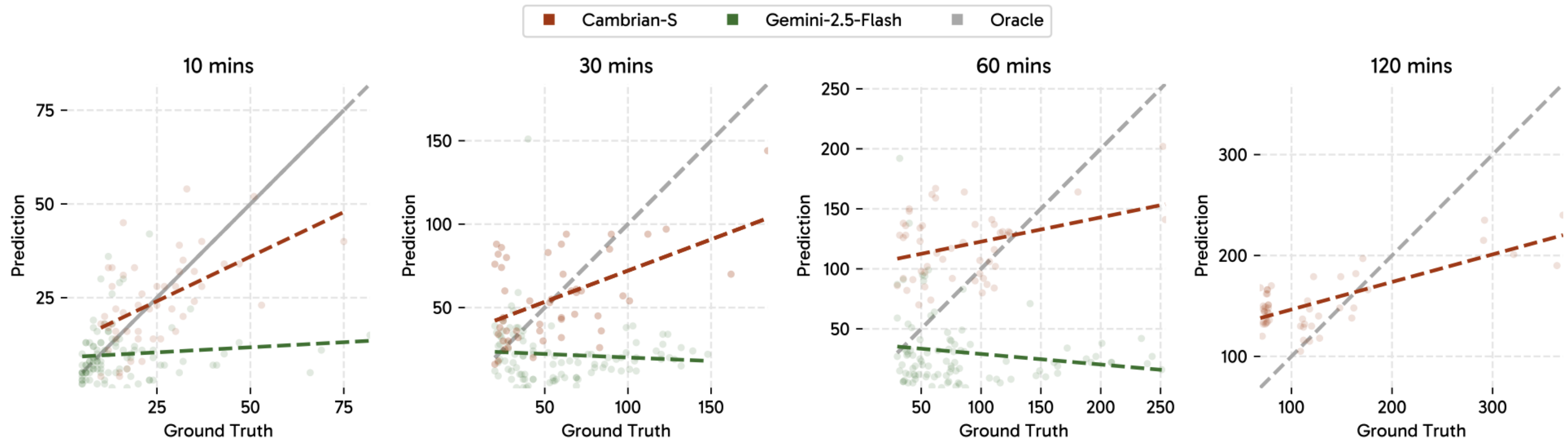



Figure 18 | **Ground truth *vs.*prediction distribution.** Compared to Gemini-2.5-Flash, Cambrian-*S* with predictive error as surprise exhibits better generalizability as the ground truth number of objects increases. The gray dashed line represents perfect prediction ($y = x$).

To summarize:

**We must build artificial supersensing before artificial superintelligence.**

We are sitting on a big opportunity here, literally.

Thank You!